Optimal Rationing within a Heterogeneous Population

Philippe Choné, Stéphane Gauthier

Optimal Rationing within a Heterogeneous Population

Forthcoming in the *Journal of Public Economic Theory*

Philippe Choné*
CREST-ENSAE

Stéphane Gauthier†
PSE and University of Paris 1

April 11, 2016

Abstract

A government agency delegates to a provider (hospital, medical gatekeeper, school, social worker) the decision to supply a service or treatment to individual recipients. The agency does not perfectly know the distribution of individual treatment costs in the population. The single-crossing property is not satisfied when the uncertainty pertains to the dispersion of the distribution. We find that the provision of service should be distorted upwards when the first-best efficient number of recipients is sufficiently high.

JEL classification numbers: I18, D82, D61

Keywords: Rationing, screening, universal coverage, upward distortion, Spence-Mirrlees condition

*ENSAE, 3 avenue Pierre Larousse, 92245 Malakoff; Phone: +33(0)141175390; E-mail address: philippe.chone@ensae.fr.

†MSE, 106-112 bd de l'Hôpital, 75013 Paris, France; Phone: +33(0)144078289; E-mail address: stephane.gauthier@univ-paris1.fr.

1 Introduction

We consider an agency in charge of supplying a service or a treatment to a population of potential recipients. Examples include medical procedures for patients, after-school programs for low-income children, social care for the disabled, or training programs for the unemployed. When the cost and the benefit of the treatment vary across individuals, efficiency recommends to supply the service only to those with a low enough cost benefit ratio. When these variables are well observed by the agency, rationing by denial can be used to implement the efficient policy.

In many instances, however, there remains substantial unobserved heterogeneity in cost and benefit conditional on observable variables. The agency then may then observe the number of treated recipients, but not their individual characteristics. In this circumstance it can rely on rationing by selection, i.e., leave the selection of recipients to the discretion of a better informed service provider (Klein and Mayblin (2012)) and fund the system on the basis of the number of treated recipients. However, because the provider's preferences are in general not perfectly aligned with those of the agency, the population of selected recipients typically departs from the first-best efficient recommendation. The literature generally assumes that the agency, when designing the provider's compensation scheme, perfectly knows the underlying distribution of cost and benefit in the population of potential recipients. For instance, the assumption is made in Makris and Siciliani (2013) and Malcomson (2005), with the former article investigating provider altruism and the latter considering many treatment varieties –two issues not addressed here.

Perfect knowledge of the cost distribution is a strong assumption. Indeed, because of data availability, only few studies have estimated the distribution of individual treatment costs in specific contexts. It is obviously difficult for researchers and agencies to obtain information about all the individual characteristics that may affect treatment costs. One first difficulty comes from a possible selection bias when only the characteristics of the treated recipients are observed. Even assuming that a researcher observes all relevant variables for a particular sample of recipients, econometric analysis only provides agencies with statistical estimates whose precision depends, among other things, on the size of the considered sample.

We argue in this article that the imperfect knowledge of the cost distribution pertains to the dispersion as well as to the mean of that distribution.

To make this point clear, it is useful to consider a typical model used by econometricians to estimate heterogenous treatment costs. Following Dormont and Milcent (2005), assume that the cost of treating recipient i by provider j at time t is given by

$$C_{ijt} = X_{ijt}\gamma + u_j + \eta_{jt} + \varepsilon_{ijt}, \tag{1}$$

where X_{ijt} are exogenous variables, u_j is a provider-specific effect that can be assumed fixed or random, and η_{jt} and ε_{ijt} are zero-mean disturbances. Assume that the linear specification (1) is correct and that all variables that influence cost are included. Suppose, furthermore, that the provider's compensation scheme is contingent on some of those variables, say X^c_{ijt}, but not on others, say X^m_{ijt}, for instance because the latter variables are not observed out of the sample used by the researcher. The assumption that the agency perfectly knows the cost distribution is debatable when no precise estimate of the parameters γ is available or when the distribution of the missing variables is unknown –a situation likely to be frequent in practice. Furthermore, the ignored variables X^m_{ijt} generate shifts in the mean as well as in the variance of the cost distribution across providers, i.e., both the conditional expectation $\mathbb{E}(X^m_{ijt}\gamma \mid j)$ and variance $\mathbb{V}(X^m_{ijt}\gamma \mid j)$ vary across providers.

In our framework, the agency, if it knew perfectly the cost distribution, would be able to implement the first-best policy through a well-designed compensation scheme, even though some relevant characteristics of potential recipients are unobserved. Otherwise the uncertainty about the distribution of heterogeneity causes the number of recipients to be distorted relative to first-best efficiency. The direction of the distortion depends on whether the uncertainty pertains to the mean or to the dispersion of individual treatment costs. In the former case, we find that the usual Spence-Mirrlees condition is satisfied and the distortion is necessarily downwards: The number of treated recipients is lower than recommended by first-best efficiency. In the latter case, the Spence-Mirrlees condition no longer holds. We find that the first-best optimum then governs the sign of distortion in the second-best program: The distortion is upwards when the first-best number of treated recipients is sufficiently high. Uncertainty about the cost dispersion pushes towards universal coverage policies.

2 Model

A population of individuals recipients, whose size is normalized to one, is eligible for a treatment supplied by a single provider. Individuals are indexed by two nonnegative real numbers b and c that may be correlated. The treatment of a type (b, c) recipient yields benefit b to the recipient and costs c to the provider. The corresponding net social benefit is $b - (1 + \lambda)c$, where λ is the (exogenously given) marginal cost of public funds.

Assumption 1. *The (expected) net social benefit of treatment for a given cost level c, $\mathbb{E}(b \mid c) - (1 + \lambda)c$, is a non-increasing function of cost c with a unique zero, denoted by c^{**}.*

Assumption 1 holds true when the expected benefit decreases with cost, a case often considered in the literature, e.g., in De Fraja (2000) and Makris and Siciliani (2013). Malcomson (2005) provides health related examples where this assumption is also relevant. Under this assumption, the first-best requires to treat recipients with cost $c \leq c^{**}$. Denoting by F the marginal distribution of individual treatment costs in the population of recipients, the first-best efficient number of treated recipients is $n^{**} = F(c^{**})$.

In this paper, we assume that the agency relies on rationing by selection: The agency observes the number n of recipients but not their individual characteristics. The treatment decision is delegated to a single provider who observes the individual characteristics of recipients. The agency offers a take-or-leave-it contract specifying the number of recipients that must be treated by the provider and a compensating transfer T. The utility of the provider when she accepts the contract is $U(n, T) = T - C(n)$, where $C(n)$ represents the aggregate cost of treating n recipients.

Given (n, T), utility maximization requires that the least costly recipients be treated in priority. The provider's cost of treating n recipients, $0 \leq n \leq 1$, is therefore given by

$$C(n) = \int_0^{F^{-1}(n)} c \, \mathrm{d}F(c).$$

The marginal cost is $C'(n) = F^{-1}(n)$, i.e., the cost of the marginal treated recipient is $F^{-1}(n)$, the nth-percentile of the distribution F. It follows that the cost function $C(n)$ is convex in n.

The net social benefit of treating n recipients is given by $S(n) = B(n) - (1+\lambda)C(n)$, where

$$B(n) = \int_0^{F^{-1}(n)} \mathbb{E}(b \mid c)\, \mathrm{d}F(c)$$

represents the (expected) gross social benefit. Under Assumption 1, the net social benefit function $S(n)$ is concave in n, reaching its maximum at $n^{**} = F(c^{**})$. When the agency knows the distribution of individual recipients characteristics, it can choose the number of treated recipients n and the transfer T to the provider that maximize

$$B(n) + U(n,T) - (1+\lambda)T = S(n) - \lambda U(n,T)$$

subject to the provider's participation constraint $U(n,T) \geq 0$. The solution to this maximization problem is to set $n = n^{**}$ and $U = U^{**} = 0$.

Under Assumption 1, the provider treats in priority the recipients with highest expected social values: The social net benefit and the provider's private objective are aligned. Hence, the agency can achieve first-best efficient solution if it knows the distribution of individual recipients characteristics. From that distribution, the agency infers the number n^{**} and the corresponding cost, $C(n^{**})$. It is then sufficient to ask the provider to treat n^{**} recipients and reimburse $C(n^{**})$.

3 Unknown cost distribution

We now relax the assumption that the marginal distribution of individual treatment costs, F, is perfectly known to the agency. We consider the simplest form of uncertainty, whereby F takes two possible values, F_H and F_L with associated probability π_H and π_L. Hence, the agency is faced with one provider, itself confronted with a population of recipients within which the marginal distribution of individual cost is either F_H or F_L. The actual marginal cost distribution function is private information to the provider. We refer to $i \in \{H, L\}$ as the provider type. Provider i has cost function $C_i(n)$ with $C_i'(n) = F_i^{-1}(n)$.

Assuming that the distribution of benefit conditional on cost is the same for both providers, we write the net social benefit of having n recipients

treated by provider i as

$$S_i(n) = \int_0^{F_i^{-1}(n)} [\,\mathbb{E}(b|c) - (1+\lambda)c\,]\,\mathrm{d}F_i(c).$$

Assumption 1 is supposed to hold for the two provider types: The first-best cost threshold for provider i is c^{**} such that $\mathbb{E}(b|c^{**}) = (1+\lambda)c^{**}$. The corresponding first-best efficient number of treated recipients is $n_i^{**} = F_i(c^{**})$.

Appealing to the revelation principle, we assume without loss of generality that the agency offers a menu (n_i, T_i), $i = H, L$, maximizing

$$\sum_i \pi_i \left[S_i(n_i) - \lambda U_i(n_i, T_i)\right]$$

subject to the provider's participation constraints $U_i(n_i, T_i) = T_i - C_i(n_i) \geq 0$ and the incentive constraints $U_i(n_i, T_i) \geq U_i(n_j, T_j)$ for all $i, j = H, L$.

Suppose first that F_L first-order stochastically dominates F_H: $F_L(c) \leq F_H(c)$ for all c, with strict inequality on a non-degenerated interval. In this case, the marginal cost functions are ordered, i.e.,

$$C_H'(n) = F_H^{-1}(n) < C_L'(n) = F_L^{-1}(n), \tag{2}$$

for all n. Given that $C_H(0) = C_L(0)$, this implies that the cost functions themselves are ordered in the same way, $C_H(n) < C_L(n)$ for all n. The first-best menu $(n_i^{**}, C_i(n_i^{**}))$ is not incentive compatible. However the usual single-crossing condition is satisfied:

$$\left.\frac{\partial T}{\partial n}\right|_{U_H} = C_H'(n) < C_L'(n) = \left.\frac{\partial T}{\partial n}\right|_{U_L},$$

for all n. This is the standard pattern in adverse selection problems. The efficient provider can mimic the inefficient one and gets the informational rent $U_H = C_L(n_L) - C_H(n_L)$. As $\pi_L S_L - \lambda\pi_H(C_L - C_H)$ decreases with n_L above n_L^{**}, the number of patients treated by the inefficient provider is distorted downwards relative to n_L^{**}.

The focus of our paper is on the case where F_H is a mean-preserving spread of F_L: Both distributions F_H and F_L have the same mean,

$$C_L(1) = \int_0^\infty c\,\mathrm{d}F_H(c) = \int_0^\infty c\,\mathrm{d}F_L(c) = C_H(1),$$

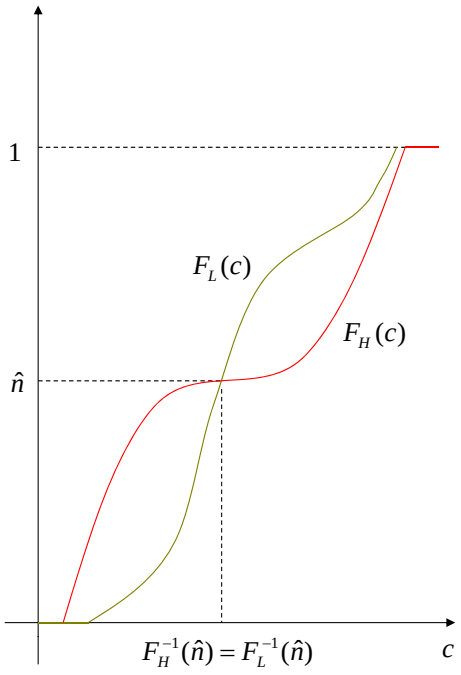

Figure 1: Individual cost cdf

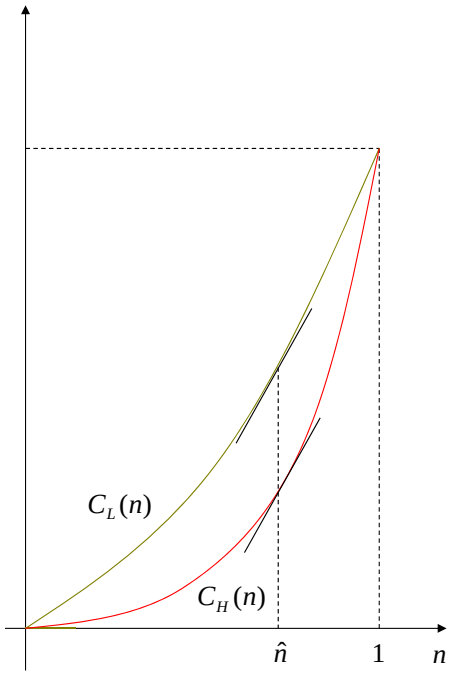

Figure 2: Aggregate cost function

and F_L second-order stochastically dominates F_H,

$$\int_0^c F_L(c)\,\mathrm{d}c \leq \int_0^c F_H(c)\,\mathrm{d}c \quad \text{for all } c.$$

Assumption 2. *The two distribution functions cross only once, at some individual treatment cost denoted* $\hat{c}$.

Under Assumption 2, we have $F_H(c) > F_L(c)$ for $c < \hat{c}$ and $F_H(c) < F_L(c)$ for $c > \hat{c}$, as shown on Figure 1. It follows that the efficient numbers of treated patients are no longer ordered as simply as under first-order stochastic dominance. Specifically, $n_L^{**} = F_L(c^{**})$ is higher than $n_H^{**} = F_H(c^{**})$ if and only if c^{**} is larger than $\hat{c}$.

Moreover, the provider marginal cost functions are no longer ordered as in (2). Setting $\hat{n} = F_L(\hat{c}) = F_H(\hat{c})$ and using the link between marginal costs and percentiles, we find $C'_H(n) < C'_L(n)$ for $n < \hat{n}$ and $C'_H(n) > C'_L(n)$ for $n > \hat{n}$. The cost difference $C_L(n) - C_H(n)$ increases from zero to $C_L(\hat{n}) - C_H(\hat{n})$ as n rises from zero to $\hat{n}$, then decreases to zero as n goes to one.

Figure 2 shows that cost dispersion translates into an efficiency advantage under rationing by selection. The efficient provider, provider H, is the one faced with the most dispersed individual treatment costs. When required to treat a given number of patients, that provider indeed has more freedom of choice when picking in priority the least costly ones in the population.

The single-crossing property does not hold, but Assumption 2 restricts the pattern of violation of this assumption, delimitating two different regions:

$$\left.\frac{\partial T}{\partial n}\right|_{U_H} = C'_H(n) < C'_L(n) = \left.\frac{\partial T}{\partial n}\right|_{U_L} \iff n < \hat{n}, \tag{3}$$

Assumption 2 yields a partitioning of the space of allocations into a positive single crossing area and a negative single crossing area similar to the one used by Araujo and Moreira (2010). The resulting single crossing in the cost difference allows us to sign local distortions in our adverse selection problem.[1]

Proposition 1. *Suppose that F_H is a mean-preserving spread of F_L and that Assumption 2 holds.*

If $c^{**} > \hat{c}$ ($c^{**} < \hat{c}$), the agency has a local incentive to distort the number of recipients treated by provider L upwards (downwards) from n_L^{**}.

Proof. At the efficient allocation (n_H^{**}, n_L^{**}), the transfers T_H and T_L that maximize the welfare are such that the efficient provider earns the informational rent $U_H = C_L(n_L^{**}) - C_H(n_L^{**}) > 0$ and the inefficient provider earns $U_L = 0$. The inefficient provider's incentive constraint can thus be written as $0 \geq C_L(n_L^{**}) - C_H(n_L^{**}) + C_H(n_H^{**}) - C_L(n_H^{**})$, which is equivalent to

$$\int_{n_H}^{n_L} [C'_L(n) - C'_H(n)] \, \mathrm{d}n \leq 0.$$

Under Assumption 2, the above inequality holds at the efficient allocation (n_H^{**}, n_L^{**}), and it is strict when $c^{**} \neq \hat{c}$. For instance, when $c^{**} > \hat{c}$, we have $n_L^{**} > n_H^{**} > \hat{n}$ and $C'_L(n) - C'_H(n) < 0$ for $n \in [n_H^{**}, n_L^{**}]$. The inefficient provider's incentive constraint is therefore satisfied in a neighborhood of the first-best allocation (n_H^{**}, n_L^{**}).

Locally, when choosing n_L close to n_L^{**}, the regulator faces a standard rent versus efficiency tradeoff, choosing n_L that maximizes

$$K(n_L) \equiv \pi_L S_L(n_L) - \lambda \pi_H [C_L(n_L) - C_H(n_L)]. \tag{4}$$

[1] In our setup, however, $C'_H(n) - C'_L(n)$ is non-monotonic in n, hence Araujo and Moreira (2010)'s Assumption A2 does not hold.

The first derivative of this function at n_L^{**} is $\lambda\pi_H[C'_H(n_L^{**}) - C'_L(n_L^{**})]$, since $S_L(n_L^{**}) = 0$. By (3), it is positive if and only if $n_L^{**} > \hat{n}$. □

Heterogeneity is about the dispersion of the marginal cost distribution implies a failure of the Spence-Mirrlees condition. This failure can be exploited by the agency to reduce the type H informational rent $C_L(n_L) - C_H(n_L)$. For n_L close to n_L^{**} the rent is decreasing in the number of recipients treated by provider L when $n_L^{**} > \hat{n}$. The agency then has a local incentive to increase the number of treated above n_L^{**}. Thus, in contrast to the standard result in the literature, the sign of the local distortion is driven in our setting by the first-best number of treated recipients: A high number of treated recipients at the first-best optimum yields a local upward distortion at the second-best optimum.

Remark. It is known that upward distortions may result from countervailing incentives arising when the type-dependence of the outside options induces the inefficient type to mimic the efficient type (see, e.g., Jullien (2000) and Lewis and Sappington (1989)). This is not the case in our setup where, due to Assumption 2, the only relevant incentive constraint is that of the efficient provider. ■

The usual single crossing assumption in the cost deals with both local and global distortions from the first-best optimum. Our single crossing assumption in the cost differences does not ensure that the global optimum involves treating $n_L^* < n_L^{**}$ recipients in the case where n_L^{**} is above $\hat{n}$. To go beyond the local results of Proposition 1, we introduce

Assumption 3. *The distributions F_L and F_H are symmetric around $\hat{c}$, i.e. $F_i(c) + F_i(2\hat{c} - c) = 1$ for all c and $i = H, L$.*

Under Assumptions 2 and 3, the distributions are equal to one half at the point where they cross: $\hat{n} = F_H(\hat{c}) = F_L(\hat{c}) = 1/2$.

Proposition 2. *Suppose that Assumptions 2 and 3 hold, F_H is a mean-preserving spread of F_L, and n_L^{**} is larger than 1/2. Then the number of recipients treated by provider L is distorted upwards if $S_L(1) > S_L(1 - n_L^{**})$.*

Proof. Assuming that provider L's incentive constraint is slack, which we check later, the second-best number n_L^* of recipients treated by that provider maximizes the function $K(n)$ given by (4). By Assumption 3, the rent $U_H(n_L) = C_L(n_L) - C_H(n_L)$ is symmetric around its global maximum at

$\hat{n} = 1/2$, i.e., $U_H(n) = U_H(1-n)$ for all n. We want to show that $n_L^* > n_L^{**}$ when $n_L^{**} > 1/2$. The proof proceeds in three steps:

1. Let $\tilde{n}_L = 1 - n_L^{**} < 1/2$. Since $U_H(n) = U_H(1-n)$ and S_L increases below n_L^{**}, the maximum of K on $[\tilde{n}_L, n_L^{**}]$ is achieved above $1/2$.

2. Since S_L is increasing and U_H is decreasing on the interval $[1/2, n_L^{**}]$, K is increasing on this interval, implying that the maximum of K on $[\tilde{n}_L, n_L^{**}]$ is achieved at n_L^{**}.

3. By symmetry of U_H and monotonicity of S_L on the intervals $[0, n_L^{**}]$ and $[n_L^{**}, 1]$, we have

$$K(n) - K(1-n) = S_L(n) - S_L(1-n) \geq S_L(1) - S_L(1-n_L^{**}) > 0$$

for all $n \geq n_L^{**}$. It follows that the maximum of $K(n)$ on $[0,1]$ is achieved at the right of n_L^{**}.

By Proposition 1, the maximum is achieved at n_L^* strictly above n_L^{**}. Provider L's incentive constraint is slack because inequality (3) evaluated at (n_H^{**}, n_L^*) is strict as $n_L^* > n_L^{**} > n_H^{**} > \hat{n}$ and $C_L' < C_H'$ on $[n_H^{**}, n_L^*]$. □

This global result is closely reminiscent of Proposition 1: The number of recipients is distorted upwards from the first-best optimum if the agency prefers the inefficient provider (the one faced with less dispersed individual treatment costs) to treat a sufficiently high number of recipients.

The global condition in Proposition 2 depends on whether treating all the population is socially preferred to treating only a fraction $1 - n_L^{**}$. It should be satisfied in practice: If the first-best optimum recommends to treat $n_L^{**} = 80\%$ of the population, the agency should prefer providing universal coverage to treating only $1 - n_L^{**} = 20\%$ of the population.[2]

References

Araujo, A., and H. Moreira (2010): "Adverse selection problems without the Spence-Mirrlees condition," *Journal of Economic Theory*, 145(3).

[2]This condition follows from the symmetry Assumption 3. Global results involving an upwards distortion obtain for asymmetric distributions when, for high values of n ($n > \hat{n}$), the first-best surplus for the inefficient provider is sufficiently high, and the difference between the aggregate cost functions of both providers decreases sufficiently with n.

De Fraja, G. (2000): "Contracts for health care and asymmetric information," *Journal of Health Economics*, 19, 663–677.

Dormont, B., and C. Milcent (2005): "How to Regulate Heterogeneous Hospitals?," *Journal of Economics & Management Strategy*, 14(3), 591–621.

Jullien, B. (2000): " Participation Constraints in Adverse Selection Models," *Journal of Economic Theor*, 93, 1–47.

Klein, R., and J. Mayblin (2012): "Thinking about rationing," Discussion paper, The King's Fund.

Lewis, T. R., and D. E. M. Sappington (1989): "Countervailing incentives in agency problems," *Journal of Economic Theory*, 49(2), 294–313.

Makris, M., and L. Siciliani (2013): "Optimal incentive schemes for altruistic providers," *Journal of Public Economic Theory*, 15(5), 675–699.

Malcomson, J. (2005): "Supplier Discretion over Provision:Theory and an Application to Medical Care," *RAND Journal of Economics*, 36(2).